Predicting Idea Co-Construction in Speech Data using Insights from Sociolinguistics

Gahgene Gweon, Mahaveer Jain, John McDonough, Bhiksha Raj, Carolyn Penstein Rose,
Carnegie Mellon University, 5000 Forbes avenue, Pittsburgh, PA 15213,
Email: {ggweon, mahaveercse, johnmcd, bhiksha, cprose}@cs.cmu.edu

Abstract: Automatic assessment of group processes in collaborative groups is one of the holy grails of the computer supported collaborative learning community. In this paper, we present work towards detecting one type of group process which provides an important window into the inner workings of a group, namely “idea co-construction (ICC)”. What is unique about our approach in relation to other educational data mining techniques is that we adopt insights from sociolinguistic theories by modeling stylistic convergence of speech. We present an unsupervised machine learning technique that is able to generate a predictor of the prevalence of ICC in face-to-face debates with an $R^2$ value of .13, $p < .05$.

Introduction

Automatic analysis of collaborative processes has value both as part of dynamic interventions in the midst of collaborative learning as well as for research towards understanding how collaboration works. Early work in collaborative learning process analysis focused on text based interactions (Rosé et al., 2008; McLaren et al., 2009; Mu et al., 2011). This work has enabled a whole series of studies where interactive support for collaborative learning was triggered by real time analysis of collaborative processes (Chaudhuri et al., 2008; Ai et al., 2010; Kumar et al., 2011). As communication technologies such as cell phones and voice over IP become more ubiquitous and allow for communication and collaboration over multiple modalities including video, audio, and text to be accessible any time and any place, the line between online group learning and face-to-face group learning begins to blur. Furthermore, as more and more collaboration takes place over video and audio channels, the need grows for the CSCL community to think about how to extend collaboration support technologies from the text realm into audio and eventually video. Recently, interest in group learning supported by robots has also begun to emerge. These shifts towards face-to-face group interactions in the three dimensional world around us rather than online require a corresponding shift in analysis technology from text-based input to speech based input. To meet this challenge, early work towards analysis of collaborative processes from speech has begun to emerge as well (Gweon et al., 2011), although the early results showed predictive value that was just above random.

Machine learning algorithms are designed with the goal of finding mappings between sets of input features and output categories. When it comes to applications of machine learning to speech, the algorithms are not applied to the speech data in its raw form. Instead, it must first be represented in terms of a list of attribute-value pairs referred to collectively as a vector space representation of the speech. Thus, first the researcher must decide what is the set of features that every segment of speech will be represented in terms of. And then for each segment, these features must be extracted so that each attribute is associated with a value that was extracted from the speech. Supervised machine learning algorithms find stable patterns within these feature vector representations by examining collections of hand-coded “training examples” for each output category, then using statistical techniques to find characteristics that exemplify each category and distinguish it from the other categories. The goal of such an algorithm is to learn general rules from these examples, which can then be applied effectively to new data. In order for this to work well, the set of input features must be sufficiently expressive, and the training examples must be representative. The contribution of this paper is the application of an unsupervised machine learning technique that is able to generate a predictor of the prevalence of ICC in face-to-face debates with an $R^2$ value of .13. The advantage of an unsupervised technique is that it can be trained on data that has not been hand labeled.

The area of automatic collaborative process analysis has focused on discussion processes associated with knowledge integration such as transactivity (Berkowitz & Gibbs 1983; Teasley 1997; Weinberger & Fischer 2006), inter subjective meaning making (Suthers 2006), and productive agency (Schwartz 1998). We refer to the discussion process that we are targeting as idea co-construction (ICC), which is similar in nature to these earlier constructs. More specifically, ICC is defined as the process of building on an idea expressed earlier in a conversation using a reasoning statement. Research has shown that such deep knowledge integration processes provide opportunities for cognitive conflict to be triggered within group interactions, which may eventually result in cognitive restructuring and learning (de Lisi & Golbeck 1999). While the value of this general class of processes in the learning sciences has largely been argued from a cognitive perspective, these processes undoubtedly have a social component. At the heart of the Piagetian roots of the concept of transactivity comes the idea that it occurs when there is a balance of assimilation and accommodation within an
interaction. We expect this balance to occur when discussion partners share a certain equality in footing. Thus, with respect to the goal of automatic analysis of ICC from speech data, we hypothesize that a feature representation that captures speech structure related to social processes that reflect effort to build this social balance into an interaction will be highly predictive of the prevalence of ICC in an interaction. We test our hypothesis on a corpus of face-to-face debate discussions collected as part of research on arousal and learning (Nokes et al., 2010).

In the remainder of the paper, we first situate our work in the midst of current directions in speech processing and review the literature on speech style accommodation in order to motivate our hypothesis. Next, we present both our manual and automatic approach for measuring the prevalence of ICC in debate discussions. We present an evaluation of the predictive value of our model. We conclude with a discussion of future directions.

**Motivation and Background**
The research goal of the work presented in this paper is to develop a speech processing technique that is capable of predicting the prevalence of idea co-construction (ICC) contributions in speech, where an ICC contribution is one in which reasoning is made explicit, and that reasoning builds on a prior reasoning statement in the discussion.

Automatic analysis of deep knowledge integration processes is not a new direction in the CSCL community in itself. For example, several researchers have applied machine learning using text, such as newsgroup style interactions (Rosé et al., 2008), chat data (Joshi & Rosé, 2007), and transcripts of whole group discussions (Ai et al., 2010). The unique contribution of the work presented here is that it is one of the first approaches to detection of ICC in speech, and the first that does so by leveraging insights from sociolinguistics. In state of the art approaches to applying machine learning technology to speech data, including the earlier work on detecting ICC from speech (Gweon et al., 2011), the speech signal is first processed using basic audio processing techniques in order to extract features from segments of speech, which are then used for classification using a machine learning model.

Gweon and colleagues (2011) used the same acoustic and prosodic features that are typically used in the language technologies community for predicting emotion from speech. For example, Ranganath and colleagues (2009) used acoustic and prosodic features extracted from speech data to predict whether a speaker came across as flirting or not in a speed dating encounter. Similarly, Ang and colleagues (2002) and Kumar and colleagues (2006) applied a similar technique to the problem of detecting emotions such as boredom, confusion, or surprise, whereas Liscombe and colleagues (2005) applied these techniques to the problem of detecting student uncertainty. All of this work makes use of signal processing tools that are able to extract basic acoustic and prosodic features such as variation and average levels of pitch, intensity of speech, amount of silence and duration of speech. What this general approach misses is the way the values of these features change over the course of an interaction and how that change itself is meaningful. It is the social interpretation of the change in speech style characteristics that is the crux of the contribution of this paper.

Acoustic and prosodic features are frequently associated with intuitive interpretations that make them an attractive choice to play a role in baseline techniques for these stylistic classification tasks. For example, increased variation in pitch might indicate that the speaker wants to deliver his ideas more clearly. Likewise, volume and duration of speech may signal that the speaker is explaining his ideas in detail, and is presenting his point of view about the subject matter. What is different about our work is that we base our approach on insights about the way in which speech style specifically (Coupland, 2009; Eckert & Rickford, 2001; Jaffe, 2009) and language style more generally (Fina et al., 2006) reflect both intentional and subconscious aspects of the way in which a speaker positions him or herself within an interaction at multiple levels. These recent accounts build on decades of work beginning with Labov’s work on speech characteristics that signal social stratification (Labov 1966) and Giles’ work developing Social Accommodation Theory (Giles 1984), which describes how speech characteristics shift within an interaction, and how these shifts are interpreted.

While the essence of transactivity has been characterized in prior work in terms of content level distinctions, we argue that it also has a social interpretation. For example, Azmitia and Montgomery (1993) have demonstrated that friends exhibit higher levels of transactive conversational moves, which are operationalized in a way that is similar to ICC, than pairs that are not friends. Furthermore, it makes sense to consider that in order to build on a partner’s reasoning, one must be attending to the partners reasoning and deem it worth referring to in the articulation of one’s own reasoning. What we argue here is that we expect a construct known as Speech Style Accommodation to reflect these social processes and thus predict prevalence of ICC.

**Defining Speech Style Accommodation:** The concept of what we refer to as Speech Style Accommodation has its roots in the field of the Social Psychology of Language. In this field, the many ways in which social processes are reflected through language, and conversely, how language influences social processes, are the
target of investigation (Giles & Coupland, 1991). As a first step towards leveraging this broad range of language processes, we refer to one very specific topic, which has been referred to as entrainment, priming, accommodation, or adaptation in other computational work (e.g., Levitan, Gravano, & Hirschberg, 2011). Specifically we refer to the finding that conversational partners may shift their speaking style within the interaction, either becoming more similar or less similar to one another.

Our usage of the term accommodation specifically refers to the process of speech style convergence within an interaction. Stylistic shifts may occur at a variety of levels of speech or language representation. For example, much of the early work on speech style accommodation focused on regional dialect variation, and specifically on aspects of pronunciation, such as the occurrence of post-vocalic r in New York City, that reflected differences in age, regional identification, and socioeconomic status (Labov, 2010). Distributions of backchannels and pauses have also been the target of prior work on accommodation (Levitan et al., 2011). These effects may be moderated by other social factors. For example, Bilous and Krauss (1988) found that females accommodated to their male partners in conversation in terms of average number of words uttered per turn. Additionally, Hecht, Boster, and LaMer (1989) reported that extroverts are more listener adaptive than introverts and hence extroverts converged more in their data.

Social interpretation of Speech Style Accommodation: It has long been established that while some speech style shifts are subconscious, speakers may also choose to adapt their way of speaking in order to achieve strategic social effects within an interaction (Sanders, 1987). One of the main motives for accommodation is to decrease social distance. On a variety of levels, speech style accommodation has been found to affect the impression that speakers give within an interaction. For example, Welkowitz and Feldstein (1970) found that when speakers become more similar to their partners, they are liked more by partners. Another study by Putman and Street (1984) demonstrated that interviewees who converge to the speaking rate and response latency of their interviewers are rated more favorably by the interviewers. Giles and colleagues (1987) found that more accommodating speakers were rated as more intelligent and supportive by their partners. Conversely, social factors in an interaction affect the extent to which speakers engage in, and some times chose not to engage in, accommodation. For example, Purcell (1984) found that Hawaiian children exhibit more convergence in interactions with peer groups that they like more. Bourhis and Giles (1976) found that Welsh speakers while answering to an English surveyor broadened their Welsh accent when their ethnic identity was challenged. Scotton (1985) found that few people hesitated to repeat lexical patterns of their partners to maintain integrity.

Computational models of speech style accommodation: Prior research has attempted to quantify accommodation computationally by measuring similarity of speech and lexical features either over full conversations or by comparing the similarity in the first half and the second half of the conversation. For example, Edlund and colleagues (2009) measure accommodation in pause and gap length using measures such as synchrony and convergence. Levitan and Colleagues (2011) found that accommodation is also found in special social events of conversation such as backchannels. They show that speakers in conversation tend to use similar kinds of speech cues such as high pitch at the end of utterance to invite a back channel from their partner. In order to measure accommodation on these cues, they compute the correlation between the numerical values of these cues used by partners.

When stylistic shifts are focused on specific linguistic features, then measuring the extent of the stylistic accommodation is simple since a speaker’s style may be represented on a one or two dimensional space, and movement can then be measured precisely within this space using simple linear functions. However, the rich sociolinguistic literature on speech style accommodation highlights a much greater variety of speech style characteristics that may be associated with social status within an interaction and may thus be beneficial to monitor for stylistic shifts. Unfortunately, within any given context, the linguistic features that have these status associations, which we refer to as indexicality, are only a small subset of the linguistic features that are being used in some way. Furthermore, which features carry this indexicality are specific to a context. Thus, separating the socially meaningful variation from variation in linguistic features occurring for other reasons can be like searching for a needle in a haystack. To meet this challenge, we measure accommodation using Dynamic Bayesian Networks (DBNs) (Jain et al., 2012). More details can be found in the methods section (Step 3.2).

Method

Our experiment setup for predicting the idea co-construction process is outline in Figure 1. In the current section, we detail our approach for steps 1, 2, 3.1 and 3.2.
1. **Data collection using speech recorders**

Our corpus was collected in a laboratory setting where pairs of participants were engaged in a debate in which they took opposing sides on a controversial topic. The specific task that the participants were asked to discuss was the cause of the decline of the Ottoman Empire, which has prompted controversy among historians. One perspective emphasizes factors internal to the Empire, while the other emphasizes external factors. Each of the participants was provided with background materials that supports the role of internal or external factors, and were asked to argue for the side that was assigned.

Participants were all male undergraduate students between the ages of 18 and 25. In prior studies, it has been shown that accommodation varies based on gender, age and familiarity between partners. Because this corpus controls for most of these factors, it is appropriate for our experiment. Furthermore, because the participants did not know each other before the debate, we can assume that if accommodation occurred, it was only during the conversation.

Each debate lasted 8 minutes. In order to collect clean speech with each student on a separate channel, each student wore a directional microphone. Nevertheless, although it is possible to clearly identify the main speaker from an audio file, crosstalk, which is the other participants’ voices, could still be heard in the background. A total of 20 sessions (40 participants) were collected and used for further analysis.

2. **Transcribing and segmenting the recorded data**

For each audio file, the eight-minute discussion sessions were transcribed and manually segmented for further analysis. A total of 40 meetings were collected, transcribed, and segmented according to the following three main rules. The resulting data contained a total of 1890 segments.

- **Independent-clause rule**: a segmentation boundary should be placed as soon as an independent clause is identified.
- **Dependent-clause rule**: a sentence that cannot stand alone should be unitized with either the preceding or following unit.
- **Analyze-from-Beginning rule**: sentences should be analyzed from the beginning of the sentence towards the end, i.e. Match the subject in the sentence with the closest verb.

3.1. **Computing the amount of the idea co construction**

When students are working on a given task or a project in a team, they receive a certain amount of information that would help them solve the problem, in the form of a task statement and training materials. We are interested in capturing instances when students display reasoning during group discussions that goes beyond what is given and displays some understanding of a causal mechanisms behind the information. Typically some causal mechanism would be referenced in a discussion of how something works or why something is the way it is. In segmenting student talk and identifying which segments display reasoning we are able to quantify amount of reasoning displayed. However, it is important to note that since what we are coding is attempts at displayed reasoning, we need to allow for displays of incorrect, incomplete, and incoherent reasoning to count as reasoning. That will necessarily be quite subjective – especially in the case of incoherent explanations. We begin by operationalizing the distinction between non-reasoning statements and reasoning statements, and then we focus on the distinction between reasoning statements that represent new directions within a conversation (i.e., externalizations) from those that build on prior contributions (i.e., ICC contributions).

One important goal in detecting the knowledge integration process is to distinguish instances when students are making their own reasoning explicit from ones that just parrot what they have heard. In our formulation, we consider the task and training materials provided during the experiment to be “given”, and we look for contributions where students go beyond that.
Operationalization step 1: Reasoning process

Our formulation of what counts as a reasoning display comes from the Weinberger and Fischer’s (2006) notion of what counts as an “epistemic unit”, where what they look for is a connection between some detail from the given task (which in their case is the object of the case study analyses their students are producing in their studies) with a theoretical concept (which in their work comes from the attribution theory framework, which the students are applying to the case studies). When they have seen enough text that they can see in it mention of a case study detail, a theoretical concept, and a connection between the two, they place a segment boundary. Occasionally, a detail from a case study is described, but not in connection with a theoretical concept. Or, a theoretical concept may be mentioned, but not tied to a case study detail. In these cases, the units of text are considered degenerate, not quite counting as an epistemic unit.

We have adapted the notion of an epistemic unit from Weinberger and Fischer (2006) because the topic of our conversations is very different in nature. The conversations that we analyzed come from a debate where two participants are asked to take opposing sides on the cause of the fall of the Ottoman Empire. As in Weinberger and Fisher’s (2006) notion of “epistemic unit”, we look for a connection between two or more concepts. We describe our operationalization in detail below. First, examine a segment of a conversation provided in Table 1. The fourth column indicates whether the given contribution contains reasoning (“R”) or no reasoning (“N”).

Table 1: Sample contribution

<table>
<thead>
<tr>
<th>Line</th>
<th>Speaker</th>
<th>Contribution</th>
<th>R/N</th>
<th>I/E</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>B</td>
<td>I think that the economic downfall of the Ottoman Empire was due to internal problems because of the first World War uh, and other civil wars going on uh, beforehand which took place over the hundreds and thousands of years that people have been in that area.</td>
<td>R</td>
<td>E</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>Um, this lead to, these wars lead to population problems.</td>
<td>R</td>
<td>I</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>Uh, people were either being killed or they couldn't farm,</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>and if you can't farm, you can't feed people</td>
<td>R</td>
<td>I</td>
</tr>
</tbody>
</table>

The simple way of thinking about what constitutes a reasoning display is that it has to communicate an expression of some causal mechanism. Often that will come in the form of an explanation, such as X because Y. However, it can be more subtle than that, for example “Russian invasion in 1914 led to a decrease in their population.” The basic premise was that a reasoning statement should reflect the process of drawing an inference or conclusion through the use of reason. Note that in the example with the Russian invasion, although there is no “because” clause, one could rephrase this in the following way, which does contain a “because” clause: “The population decreased because of the Russian invasion in 1914.”

More generally, we defined reasoning display as a relationship between two or more concepts. A concept could be prior knowledge, or one of the facts provided to the participants. The presence of multiple concepts in a statement by itself does not determine whether a statement contains reasoning. Rather, the relationship between multiple concepts is the determining factor. For example, a simple list of concepts (e.g., population decreased) is information sharing, and not reasoning. We identified two types of relationships that signal a reasoning process; 1. Compare & contrast, 2. Cause & effect.

1. Compare and contrast, tradeoff: When the speaker compares two concepts, the speaker is making a judgment, which involves thinking about how two concepts are related to another.
   • The speaker compares two time periods (“at the time” & “today”): “At the time if you look at the technology, it wasn’t that advanced as we have today.”
   • When a speaker makes an analogy, he is making a link due to the similarity between two concepts. “Outside powers were like the match lighting the fire.”

2. Cause and effect: When the speaker uses a cause-and-effect relationship, this process involves establishing the relationship between two concepts through a reasoning process. The general relation in this category is “doing x helps you achieve y”. Examples are illustrated below.
   • A because of B: “they forced the Empire to be economically dependent because they set up trading posts and banks”
   • A in order to achieve B: “Great Britain came in and introduced capitulations to control schools and health systems.”
Operationalization step 2: Idea co-construction (ICC) vs. Externalization

Statements that display reasoning can be either Externalizations, which represent a new direction in the conversation, not building on prior contributions, or ICC contributions, which operate on or build on prior contributions. In our distinction between Externalizations and ICC contributions, we have attempted to take an intuitive approach by determining whether a contribution refers linguistically in some way to a prior statement, such as through the use of a pronoun or deictic expression.

Take the sample conversation we used earlier to illustrate the reasoning contribution in Table 1. The last column of Table 1 is marked as either an externalization (E), idea co-construction (I) for the statements that are marked as (R). The first statement by B is an externalization since B starts a new topic, thus this contribution is not building on a prior contribution. Subsequent reasoning contributions in this discussion are coded as (I) because they each build on statements that directly precede them.

Reliability of Annotation

Two coders were initially trained using a manual that describes the above operationalization of reasoning displays and ICC in detail along with an extensive set of examples. After each coding session, the coders discussed disagreements and refined the manual as needed. Most of the disagreements were due to the interpretation of what the students meant rather than the definition of reasoning itself. Therefore, later efforts focused more on defining how much context of a statement could be brought to bear on the interpretation and how. In a final evaluation of reliability for reasoning processing, we calculated kappa agreement of 0.72 between two coders over all the data. After calculation of the kappa, disagreements were settled by discussion between the two coders. For detecting instances of ICC and externalization, our coding yielded a kappa value of 0.7.

Amount of Idea co-construction (ICC)

Because the accommodation scores were computed for each 8 minute session, we computed a comparable score by adding up the number of idea co-construction for each session. This resulted in an average score of 36. The minimum score for a session was 22, and the maximum score was 60.

3.2. Computing the amount of accommodation

What we evaluate in our study is the ability of this measurement of speech style accommodation to predict prevalence of ICC in debate conversations. In order to measure the amount of accommodation, we used Dynamic Bayesian Networks (DBNs). What we are specifically interested in is the manner in which the influence of one partner on the other is modeled. What is novel in our approach is the introduction of the concept of an accommodation state. The accommodation state models structurally the insight that accommodation occurs over time as a reflection of a social process, and thus has some consistency in the nature of the accommodation within some span of time. The major advantage of modeling consistency of motion within the style shift over time is that it provides a signpost for identifying which style variation within the speech is salient with respect to social interpretation within a specific interaction. Therefore, the model may remain agnostic to which style features are shifting and may thus be applied to a variety of interactions that differ with respect to which stylistic features have strategic social value for the participants. The technical details of the inner workings of the model as well as a validation that it is able to measure speech style accommodation are described in a separate paper (Jain et al., 2012). However, the types of features used are briefly described in the following subsection.

Speech Features

Speech stylistic information is reflected in prosodic features such as pitch, energy, and speaking rate. In our work, we leverage on several of these prosodic features to quantify accommodation using Dynamic Bayesian Networks (DBNs). To extract features, speech from each session was segmented into a window of 50ms, with adjacent overlapping windows by 40ms. From each window, a total of 7 features were computed using OPENSmile toolkit (openSmile 2011). These features are voice probability, harmonic to noise ratio, voice quality, and three different measurement for pitch ($F_0$, $F_0^{\text{raw}}$, $F_0^{\text{env}}$), and loudness. Frame size within each window was set to 50 milliseconds and frame step was set to 10 milliseconds. Next a 10 bin histogram of feature values were computed for each of these features, which was then normalized to sum to 1.0. The normalized histogram represents both the values and the fluctuation in the feature. For example, a histogram of loudness feature captures the variations in the loudness of the speaker within the turn.

Amount of Accommodation

Possible range of accommodation score ranged from 0 to 1, where 0 means there was no accommodation between the two partners. The average accommodation score was 0.43.
Results

Our hypothesis motivated by the social interpretation of ICC discussed above that we will see a significant positive correlation between the prevalence of ICC in a conversation based on our hand coding of the transcript and the amount of Accommodation detected in the speech using our Dynamic Bayesian Network model. Over the 20 discussion transcripts discussed above, we counted the number of contributions that were coded as transactive, and labeled this sum Prevalence of ICC. Similarly, we labeled the score computed by our network model as Accommodation score. When we computed a linear regression model between the Accommodation score and Prevalence of ICC, we found a statistically significant correlation, $R^2 = .13$, $p < .05$.

Conclusions and Current Directions

In this paper, we presented our work towards automatic detection of prevalence of idea co-construction contributions in speech data. The goal of this paper was to develop technology to address such needs. To this end, we have demonstrated the feasibility of using data mining techniques by modelling stylistic convergence of speech. Our work shows promise in that our model generated a predictor of amount of idea co-construction with an $R^2$ value of 0.13.

In our future work, we plan to compute the amount of accommodation at the “clause” level, the unit of analysis used for idea co-construction coding. With this level of analysis, we would have enough data points to run machine learning in order to predict whether each clause is an instance of displaying idea co-construction rather than just predicting the prevalence of ICC over the whole interaction. In addition, more sophisticated adaptations of sociolinguistic work might suggest follow-up techniques. For example, we plan to investigate sequencing and timing rather than just quantity, in line with work conducted by Kapur and colleagues (2009).

In terms of data, we are currently collecting and annotating audio data from additional meetings as well as other contexts to validate our result further as well as testing its generality across a wider variety of student groups.

References


Pon-Barry., et al. (2006). Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems. IJAIED 16(2)


Acknowledgments
This research was supported in part by NSF grants 0648487 and 0935127.